Role-Playing Game for Studying User Behaviors in Security: A Case Study on Email Secrecy

Kui Xu, Danfeng (Daphne) Yao, Manuel A. Pérez-Quiñones, Casey Link
Department of Computer Science
Virginia Tech
Email: {xmenxk, danfeng, perez}@cs.vt.edu
ctlink@vt.edu

E. Scott Geller
Department of Psychology
Center for Applied Behavior Systems
Virginia Tech
Email: esgeller@vt.edu

***Abstract*—Understanding the capabilities of adversaries (e.g., how much the adversary knows about a target) is important for building strong security defenses. Computing an adversary's knowledge about a target requires new modeling techniques and experimental methods. Our work describes a quantitative analysis technique for modeling an adversary's knowledge about private information at workplace. Our technical enabler is a new emulation environment for conducting user experiments on attack behaviors. We develop a role-playing cyber game for our evaluation, where the participants take on the adversary role to launch ID theft attacks by answering challenge questions about a target. We measure an adversary's knowledge based on how well he or she answers the authentication questions about a target. We present our empirical modeling results based on the data collected from a total of 36 users.**

I. Introduction

The ability to realistically model how much the attackers know about a target is useful. It helps predict privacy and security threats from known or unknown adversaries, which in turn facilitates the protection of confidential information. Specifically, it is desirable for one, say T, to analyze how much others including T's friends know about T's personal data, i.e., T asks "How much do others know about me?". To describe this problem more formally, given the target T, an adversary A, the history of interactions between T and A, and a sensitive piece of information $d \in \mathcal{P}$ about T from a finite space $\mathcal{P}$, we define *guessability* as the likelihood of adversary A knowing d about the target T. Solving this problem can help one model and assess security and privacy threats.

This issue – referred to by us as the *adversary's knowledge* problem – has not been addressed in the literature. There are studies on new knowledge that an adversary may gain about a target by inferring from publicly available data [1] or from online social networks [2]. In data publishing privacy, substantial amount of research has been on modeling and sanitizing data according to a varying degree of adversaries' knowledge [3], [4], [5], [6], [7], [8]. However, these solutions are not designed to address the guessability problem.

In our work, we measure an adversary's knowledge by how well he or she answers the authentication questions about a target. We quantitatively analyze factors that affect adversary's knowledge with respect to a sensitive attribute. These factors include *i)* properties of the interaction and relation between the adversary and the target directly or indirectly via third parties, *ii)* properties of the sensitive attribute, and *iii)* any public available information regarding the target. Our experimental evaluation is performed in the context of a question-based authentication system, where we evaluate one's ability to answer the challenge questions of others.

There are many types of adversaries. An adversary may be a stranger, an acquaintance, a colleague, a relative, or a close friend of a target. The adversary may be a hardened career criminal, a novice hacker, a disgruntled employee, or a cyber spy. The privacy threat and analysis may be customized under different adversary models. Without loss of generality, we present our design, model, and evaluation under a university environment. Our work analyzes the privacy threat posed by known acquaintances of a target. Our methodology applies to the analysis of other adversary models.

For our experiments, we develop a new role-playing game system that is a technical enabler for realizing our goals. The game system automatically generates challenge questions from a target's private activities. Players of the game system are asked to impersonate the target by answering the questions related to the target. This role-playing game provides a testbed for studying attack behaviors in the cyberspace.

In our user study, we collected 1,536 user responses and associated 3,072 behavior data points from experiments. Our results reveal a 41.4% average success rate when a player is asked to answer the multiple choice privacy questions of a target in a university setting. We found that the duration of relation and communication frequency between the target and the player are strong predictors.

The private information in our game system is based on a target's email messages. Email messages are usually accessible only by the owner, and thus it is reasonable to consider them as private between the sender and the receiver. We automatically generate challenge questions based on email contacts, subjects, or contents. Our experiments measure how well others know about the email activities of a target. All email messages contributed by participants are properly sanitized by their owners to remove possible sensitive information.

Our analysis is based on the data from 36 participants in our experiment, which might affect the accuracy of experimental findings. Conducting user studies or experiments involving

This work has been supported in part by NSF grant CAREER CNS-0953638, ONR grant N00014-13-1-0016, ARO YIP W911NF-14-1-0535, ICTAS and HUME Center of Virginia Tech.

COLLABORATECOM 2014, October 22-25, Miami, United States

DOI 10.4108/icst.collaboratecom.2014.257242

private and sensitive information has always been challenging. Despite the relatively small sample size, our work is the first step towards addressing the important problem of quantitative modeling of adversary's knowledge and our methodology based on the role-playing game is new.

II. Related Work

Existing research on understanding offensive behaviors in cyberspace is typically conducted through surveys, for example, on cyber-bullying [9] and on the likelihood of self-reporting crimes [10]. Scam victims' behaviors were analyzed in [11], where the scams studied are mostly from the physical worlds. In comparison, we design a role-playing attack game for analyzing cyber-security behaviors.

Currently, security-related games are mainly designed for education purposes, including one based on the popular multi-player online game Second Life [12]. We use game systems to conduct research relevant to cyber security. Our systems can also be used to educate users about important cyber-security concepts.

The security of authentication questions is also experimentally measured in the work described in [13]. Although with different goals, as a comparison, the experiment in [13] revealed that acquaintances with whom participants reported being unwilling to share their webmail passwords were able to guess 17% of their answers. And those who were trusted by their partners were able to guess their partners' answers 28% of the time. The numbers are lower than what we get using questions in the form of multiple choice questions.

The increasing use of online social networks also causes privacy issues, and sensitive information is usually either publicly provided or uploaded by other people or friends [14], [15]. Authors in [1] showed that, with a small piece of seed information, attackers can search local database or query web search engine, to launch re-identification attacks and cross-database aggregation. Their simulated result shows that large portions of users with online presence are very identifiable. The work in [16] used a leakage measurement to quantify the information available online about a given user. By crawling and aggregating data from popular social networks, the analysis showed a high percentage of privacy leakage from the online social footprints, and discussed the susceptibility to attacks on physical identification and password recovery. Using social networks as a side-channel, the authors in [17] are able to deanonymize location traces. The contact graph identifying meetings between anonymized users can be structurally correlated with a social network graph, and thereby identifying 80% of anonymized users precisely. In comparison, our work studies the privacy leak within an organization.

In personal information management, the work in [18] used a memory questionnaire to study what people remember about their email. They found out that the most salient attributes were the topic of the message and the reason for the email. People demonstrated good abilities to refind their messages in email. In the majority of tasks, they remembered multiple attributes. These findings help support our approach to use email (or other personal information) as a source of information for generating authentication questions.

Shannon's entropy [19], [20], [21] has been widely used in many disciplines, such as sensor networks [22], cryptography [23], and preference-based authentication [24]. Our quantifying activity fundamentally differs from the analysis by Jakobbson, Yang, and Wetzel on quantifying preferences [24], because of the diversity and dynamic-nature of personal activities in our model. Unlike [24], email-based challenges do not require users' to pre-select questions and setup answers.

Our work is different from the existing work [25] that uses entropy for quantifying knowledge-based authentication, in terms of goals and approaches. For example, Chen and Liginlal proposed a Bayesian network model for aggregating user's responses of multiple authentication challenges to infer the final authentication decision [25]. They also described a method for strategically selecting features (or attributes) for authentication with entropy [26]. Both pieces of work were validated with simulated data. Our work aims to predict the guessability with respect of an attacker's prior knowledge. We perform experimental validation with real-world data.

There have been continuous research advances in the field of authentication systems and their usability [27]. Our work is not to propose a new authentication method, rather we develop a general methodology for modeling adversary's knowledge. Authentication is used as an experimental evaluation tool to demonstrate our approach. There exist many research solutions on new authentication systems and their security evaluation (e.g., [28], [29], [30], [31], [32]). A conventional question-based authentication is usually used as a secondary authentication mechanism in a web site, when the user tries to reset a forgotten password. We adopt the email-based challenges proposed in [33], which conveniently allows us to perform accurate and specialized data collection, categorization, and quantitative measures on the data and attributes.

Similar to our work where email activities are used to generate challenge questions and evaluate adversary knowledge, applying user activities for security purposes has been researched in previous work [34], [35], [36]. User behaviors have been used for detecting illegal file downloads [34], discovering abnormal network traffic [35], and identifying malicious mobile apps [36].

III. System Design

We design a role-playing game system to provide a controlled and monitored environment for the players to perform the impersonation attacks against targets. We describe our design and implementation of the game system in this section. Using this system, our user study in Section V measures the guessability of personal and work email records of targets by known or unknown individuals. These individuals play the role of adversaries in this emulated ID theft scenarios in the user study.

A. Overview

We define a *target* T as the individual whose identity is being attacked, that is, a player whose challenge questions are guessed by adversaries A. A *player* aims to impersonate the target through answering or guessing the challenges. The player may know the target or may be a complete stranger to the target. The player is referred to us as the adversary.

Our evaluation can utilize any question-based authentication system. Conventional authentication questions are usually based on historic personal data and events (e.g., names of hometown and school). However, we choose not to use these conventional challenges due to two reasons, privacy and scalability. First, these types of sensitive data are used in the real world for secondary authentication; revealing it during experimental evaluation compromises the privacy of participants. Second, collecting personal data of participants requires manual efforts, which is not scalable.

Our challenge questions are generated from email messages of targets. Using emails as the data source of private information offers several advantages.

1) Email activities are dynamic and change with time, which fundamentally differ from personal facts such as mother's maiden name. Email allows us to evaluate the impact of the dynamic private data on adversaries' knowledge.
2) From a system designer's perspective, an email system allows us to completely automate operations of data retrieval, attribute extraction, challenge questions generation, and verification of user responses. We write client-side scripts utilizing email server APIs for these tasks. Email servers and email messages share the communication protocols, APIs, and data formats, which adds to the compatibility and scalability.
3) One-to-one email communication is private and suitable for our privacy evaluation. It provides a rich context and semantics for personal information. The information is not used by online commercial systems for real-world authentication.

The game system has the following components: *i)* email retrieval for retrieving email messages of targets, *ii)* question generation for parsing email messages and generating multiple-choice questions, *iii)* user interface, *iv)* web hosting for online participation and *v)* database storage for storing users' responses. Our game rules allow adversaries to search the Internet for clues and hints. Using email activities for challenge questions is desirable because of its rich context and archival nature. Our design generates email-based questions by leveraging the existing stored data of a user on the mail server.

Our design minimizes the interaction between the game server and the mail servers. We perform a one-time data transfer operation to fetch mail records of targets with proper permission and data sanitization. The corpus data is stored and analyzed by us securely for generating challenges and verifying answers. There is no subsequent interaction with the mail server. In this one-time data-transfer operation, we collect mail records, including Inbox, Sent, and local folders. Only during this data transfer, the participating target is required to enter his or her password to access the mail records on the mail server. We use JavaMail for fetching and parsing email messages. Parsing the emails allows us to extract the information such as sender/receiver, email title, timestamp and also email message data. The class IMAPSSLStore is used, which provides access to an IMAP message store over SSL. (The game server is different from the email server.)

B. Challenge Questions

We automatically generate four types of challenge questions asking about various attributes of a target's email messages. Examples are shown below.

- *FromWhom*: From whom did Professor A receive the email with subject 'Agenda for Dr. X's visit.' on 2011-03-16T14:59?
- *SentWhom*: To whom did Professor B send the email on 2011-08-18T21:21 with subject 'Re: GraceHopper 2011'?
- *FromSubject*: What is the subject of the email to Professor C from Y on 2011-06-17T13:23?
- *SentSubject*: What is the subject of the email Professor D sent to Z on Wed, Oct 5, 2011 at 5:10 PM?

A challenge question is asked in the form of multiple choices with 5 choices. Questions have wrong answers in the choices. Wrong choices are automatically generated from random email messages of the target. A question may contain a *None of the above.* choice with a pre-defined probability.

C. Overview of Game Procedure

A player logs in our server with a password through a secure HTTPS connection. Our game server hosts the challenge questions. [1]

The player selects targets to attack and answers a total of 48 challenge questions. The questions associated with the selected target are retrieved from our backend MySQL database and shown to the player in a browser. All the questions are in the form of multiple choice questions.

During the game, the player is allowed to use Internet. Upon submission, the player's answers are stored by the server. The server compares the submitted answers with the correct answers stored in the database, and computes the player's performance.

IV. Sources of Adversary's Knowledge

We categorize the factors that contribute to the leak of private information (e.g., entropy of the corresponding random variables, social relation, and interaction). We then design quantitative measurements for each of these factors, and compute their significance in predicting an adversary's knowledge.

Public information available from the Internet and public records is a common source for gaining knowledge about a target. How much knowledge about a target can be gained merely from the publicly available information on the Internet was analyzed by Yang *et al* in [1]. That study is particularly suitable for analyzing background knowledge of stranger adversaries. In contrast, our work is focused on two other factors contributing to the guessability analysis, namely data regularity, and interaction, which are described next. These factors may not be independent of each other.

- *Data regularity:* the regularity or predictability of the target's activities, profiles, or persona. This factor is

[1] Our implementation is based on Restlet Java web server.

determined by the characteristics of the target and the attribute being challenged. This factor is related to the difficulty of the challenge question. We define an activity or event to have one or more attributes describing properties of the activity. We view an attribute as a random variable that may take a number of possible outcomes. An activity may be Alice sending an email message, and its attributes may include sender, receiver, timestamp, subject of the email, and attachment of the email.
A regular event or a regular activity (e.g., the dinner location is usually at one's home) is easier to guess than a frequently changing event (e.g., the last person to whom you sent an email). We use entropy to summarize the regularity of events in our evaluation.

- *Direct or indirect relation and interaction:* the interaction and relation between the parties and their personal or workplace social network. This factor aims at capturing the dynamics between the parties in order to analyze the flow of private information. For a stranger adversary, this factor may provide no information in the analysis due to the lack of available data.
The target and the adversary may have direct or indirect social connections, so their relation and communication are important factors that can be used to estimate the knowledge of an adversary about the target. If the adversary is from the target's personal or professional social networks (e.g., relatives, colleagues, friends), the adversary has background knowledge about the target, which makes guessing easier.
A factor in modeling the adversary's knowledge is the social relations and interactions between the adversary and the target. The relation and interaction may be direct or indirect through third parties. We hypothesize that close individuals or two individuals with overlapping social networks may indicate a high degree of background knowledge about each other.
This interaction factor may be further categorized into two classes: *i)* static social relation and *ii)* dynamic interaction. The former refers to relations such as advisor-advisee, instructor-student, parent-child, friend, or colleagues. For each relationship, the dynamic interaction (e.g., duration of relation, communication patterns) between the involved parties provide more fine-grained information and description for our analysis.
To completely gather these social interactions is challenging, if not impossible, e.g., water cooler conversations are difficult to systematically record and analyze. For our experimental demonstration, we choose to analyze email records because of its archival nature.

- *Collusion among adversaries:* the collusion among adversaries is the case that multiple adversaries collaborate in figuring out one same target's private information. The share of knowledge has a big impact in the total amount of information adversaries can obtain by teaming up with each other. Different people know the target from different aspects, and by putting knowledge together, adversaries have a more complete understanding about the target, both direct and indirect.

There are various methods for quantifying these factors and integrating them to assess the adversary's knowledge. We perform regression analysis based on our quantified factor values. The resulting model can be used to assess the knowledge of either a specific individual or types of individuals.

Our results shown in Section V found that the duration of relation and frequency of communication are strong predictors of adversary's guessability in our model. These factors may be integrated with the public information factor during the analysis. The accuracy of modeling may highly depend on the completeness and accuracy of the information used in the analysis.

V. Experimental Evaluation

All our experiments involving human subjects have been conducted under proper IRB approvals and are compliant to IRB regulations. We gave extra caution to protect the data security. There are two roles in our experiment: *target* from whose email messages questions are generated, and *player* (i. e., attacker) who guesses the questions from the target. The player is allowed to use the Internet. Targets are all professors in a university. They contributed their sanitized email content through an automatic procedure. We assume that email messages are private between the sender and receiver, and contain personal and work-related information.

A. Experimental Setup

We generate 24 challenge questions from each target's email records. The questions are sanitized by the target. 12 questions are based on the sender or receiver (referred to as *SentWhom* and *FromWhom*). 12 questions are based on email subjects (referred to as *SentSubject* and *FromSubject*). We only process email headers, and the content of email messages is not kept or used.

Email header can be considered as the abstract of an email message and contains different kinds of private information which is not limited to the form of emails. It also allows easy and automatic information processing for experimental question generation. Richer information can be extracted from email contents, with advanced natural language processing and more strict sanitization. Our experimental approach can be generalized to use other sources of personal information as well.

We consider a stronger adversary model compared to complete strangers acting as attackers (e.g., as in the analysis done in [1]). The attackers could be acquaintances of their targets. To simulate such situation, we recruited students of the targets as players, including undergraduate and graduate students within the same university. Some of the students may or may have worked with the targets, so the adversaries (players) in our model may have more access to their target for gaining knowledge about the challenge questions.

It's possible that the adversary may be partly involved in some email messages with the target. However, the chance is low considering the total number of email messages each target has. Some targets provide the email messages in the Inbox or

Sent folder for experiment, while others choose to provide the email messages in a few organized folders, so the timespan of the messages collected from each target varis, from months to years.

We give players performance-based incentive cash rewards, i.e., the amount of their rewards depends on the number of correct answers. Each player answers questions about two targets (48 total). We also collect and analyze behavioral data. The behavior data includes *i)* the duration of knowing the target and *ii)* the player's confidence about his or her answer. Table I summarizes the experimental setup.

TABLE I. SUMMARY OF EXPERIMENTAL SETUP.

Target	Player	Auth. question	Behavior question
4	32	1,536	3,072

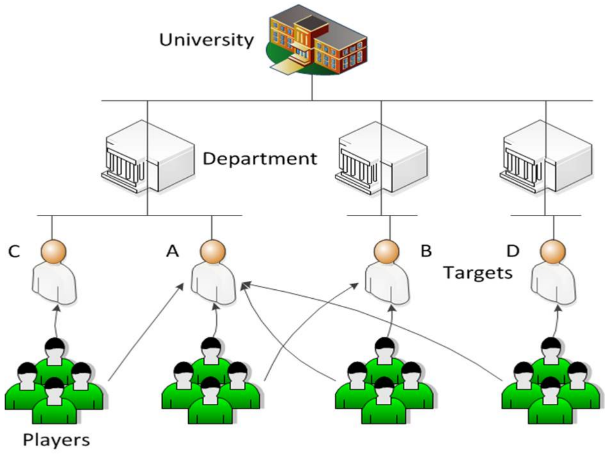


Fig. 1. Social network structure.

Figure 1 shows the target-player relations within the university. Our analysis aims to answer the following questions:

- What is the guessability of the target's questions for players with different types of relation?
- Which are the factors in terms of predicting the adversary's knowledge in a multiple regression analysis?
- How does the collusion among adversaries affect the total knowledge obtained about the targets?

B. Predictors in Regression Analysis

We analyze the factors that may contribute to the leak of personal information in our game model, including social relations and social networks. We compute the correlation between these factors and our data (namely the corresponding number of correctly guessed questions) in a regression analysis.

We summarize the explanatory variables (or predictors) in our regression analysis. Among these factors, factor IV belongs to the data regularity category, and factors I, II, III belong to the relation and interaction.

- Factor I: Type of relation between the target and the player
- Factor II: Duration of relation between the target and the player
- Factor III: Communication frequency between the target and the player
- Factor IV: Entropy of target's email regularity, specifically based on the probability distribution of the target's email frequency to and from the target's contacts.

TABLE II. NUMBER OF CORRECT GUESSES VS. TYPE OF RELATIONS.

Relation Target	Within University	Within Department	Undergrad Advisor	Graduate Advisor
	Avg. No. of Correct Answers (Sample Size)			
Prof. A	5.1 (21)	5.6 (5)	12.0 (1)	10.6 (5)
Prof. B	6.5 (6)	NA	NA	9.8 (6)
Prof. C	NA	NA	NA	8.0 (5)
Prof. D	NA	NA	7.5 (14)	17.0 (1)
Avg. Correct	5.4	5.6	7.8	9.9
Std. Error	0.4	0.9	0.7	0.7
Correct %	22.5%	23.3%	32.5%	41.4%

Factor I: type of relation. In our experiment, the type of relation between the target-player pair includes:

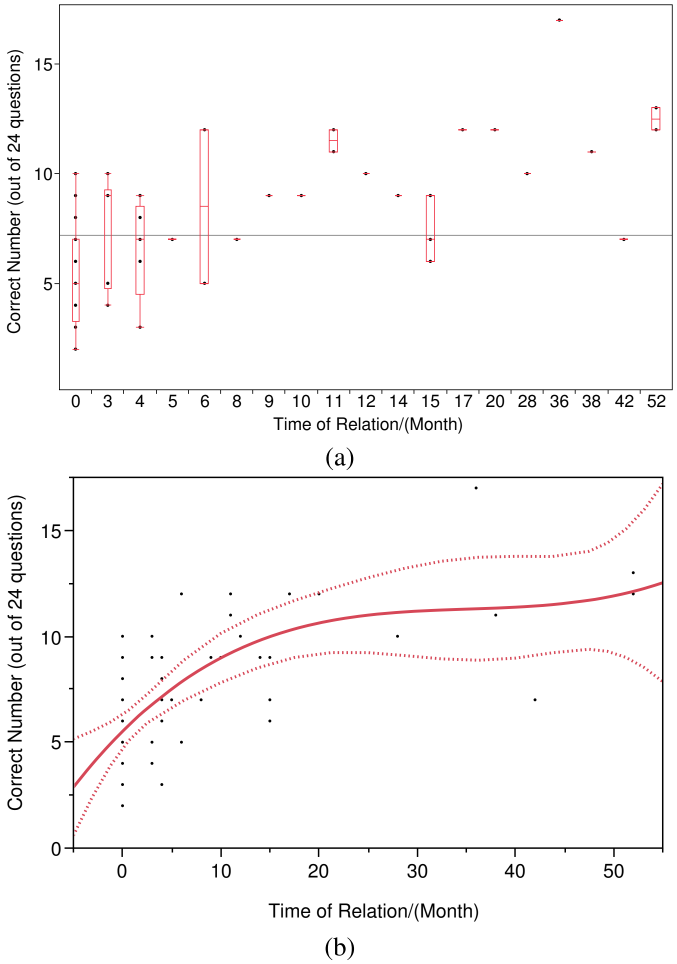
- Graduate Advisor (Grad): The target is the player's graduate advisor.
- Undergraduate Advisor (UGrad): The target is the player's undergraduate advisor.
- Within Department (WD): The target is not the player's advisor, but is in the same department with the player.
- Within University (WU): The target is neither the player's advisor, nor is in the same department, but is in the same university.

Table II gives the average numbers of correctly answered questions out of 24 for each of the four relations.

- (WU) When the target and player are a professor and a student within the same university, the player is able to guess 22.5% of the 24 questions. This result is only slightly higher than 20% for random guesses.
- (WD) The result gets slightly better at 23.3%, when the two are both in the same department.
- (UGrad) When the target is the player's undergraduate advisor, the player is able to get an average correct percentage at 32.5%.
- (Grad) The player's performance is high, when the target is his or her graduate advisor. On average, 41.4% of questions can be correctly answered.

Factor II: duration of relation. For each target-player pair, we analyze the correlation of player's performance with the duration of knowing the target. In Figure 2, Y-axis is the number of correct guesses, and the data points are grouped by their corresponding duration of relation between the target and player. The duration of relation in the X-axis is measured by month. Despite the data variance, the figure shows the positive correlation between the duration and performance as expected.

Factor III: communication frequency between the target and the player. We analyze the frequency of email



(a)

(b)

Fig. 2. Number of correct guesses for each player vs. duration of the player knowing the target. (a) A box plot shows the sample minimum, lower quartile, median, upper quartile and sample maximum of grouped data. The horizontal line in the figure is the average correct answers. (b) A polynomial fit of the data with degree 3. The dot lines show the limits for the expected value (mean), at confidence level of 95%.

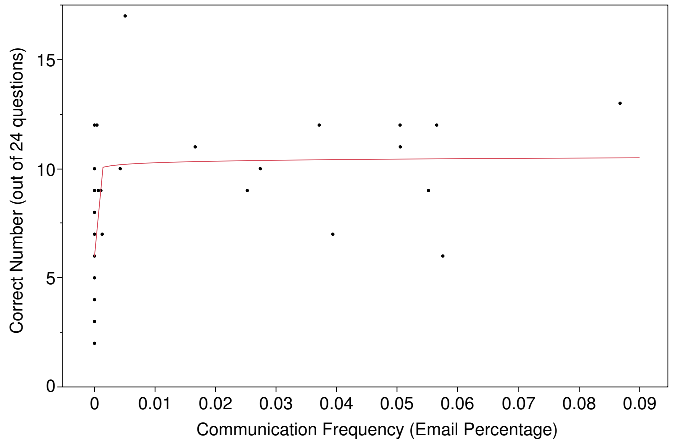


Fig. 3. Number of correct guesses for each player vs. communication frequency between the player and the target.

communication for each target-player pair. In Figure 3, each point represents a player; X-axis shows the fraction of the target's email communications that involve the player and Y-axis shows the player's performance. The result shows that having direct communication, even not very frequent, is a good indicator of more knowledge about the target. The solid line in Figure 3 shows a logarithmic fit function $Y = c_0 logX + c_1$.

Factor IV: Entropy of target's email regularity. The entropy of a target's email activities summarizes the diversity in the target's communication. The entropy value is computed from the probabilities of the contacts out of all email messages. All the email data collected from each target at the beginning of the experiment is considered as the observed data set. For example, target A has contacts c_1, c_2, c_3, and the number of emails between target A and c_1, c_2, c_3 are n_1, n_2, n_3. We compute the following probabilities for each contact:

$$\begin{cases} p_1 = n_1/(n_1 + n_2 + n_3) \\ p_2 = n_2/(n_1 + n_2 + n_3) \\ p_3 = n_3/(n_1 + n_2 + n_3) \end{cases}$$

And the entropy H_A of target A is

$$H_A = -(p_1 ln(p_1) + p_2 ln(p_2) + p_3 ln(p_3)) \quad (1)$$

C. Regression Analysis

In preparation for the regression analysis, we quantify each of the four explanatory variables T, D, C and E, and then use them to compute the guessability. T approximates the impact of relation type on the adversary's knowledge using a non-linear function.

- T: type of relation (factor I). To quantitatively represent a relation type, an exponentially decreasing function e^{-x} is used. We assign higher numerical values to more closer relation types. The choices for the function and numerical values are empirical and based on initial data analysis.

$$T = e^{-x} \begin{cases} 1, & x = 0 \text{ for Grad-relation} \\ e^{-1}, & x = 1 \text{ for UGrad-relation} \\ e^{-2}, & x = 2 \text{ for WD-relation} \\ e^{-3}, & x = 3 \text{ for WU-relation} \end{cases}$$

- D: duration of relation in months (factor II).
- C: communication frequency (factor III). We apply a logarithmic transformation to the original data.
- E: entropy of target's activity (factor IV).

Our multiple linear regression has two tasks.

1) *Task I* is on analyzing the relationship between the four explanatory variables (T, D, C, and E) and dependent variable Y (player's performance), where Y is the number of correct answers out of 24 questions for each player-target pair.
2) *Task II* is a more fine-grained analysis on the relationship between the explanatory variables and Y' (categorized player's performance), where Y' is the number of correct answers for each of the four categories of questions. The four categories of questions are *FromWhom*, *FromSubject*, *SentWhom*, and *SentSubject*.

The multiple linear regression produces a prediction function in the form of:

$$Y = f(T, D, C, E) = c_0T + c_1D + c_2logC + c_3E + c_4 \tag{2}$$

where Y is the predicted number of correctly answered questions, or the amount of confidence for a player when answering a target's questions.

For *Task I*, our regression analysis gives the following prediction function. The coefficients and their standard errors are given in Table III.

$$Y = 1.12T + 0.08D + 0.13logC + 1.09E + 2.51 \tag{3}$$

TABLE III. REGRESSION ANALYSIS RESULTS. P-VALUE INDICATES HOW MUCH EACH EXPLANATORY VARIABLE IS CORRELATED TO THE DEPENDENT VARIABLE Y. SMALLER P-VALUE MEANS HIGHER SIGNIFICANCE OF THE EXPLANATORY VARIABLE.

	Coefficients	Std Error	t Stat	P
Intercept	2.51	4.59	0.55	0.59
Relation Type	1.12	1.40	0.80	0.43
Duration	0.08	0.04	2.33	0.02
Communication	0.13	0.06	2.21	0.03
Entropy	1.09	0.83	1.31	0.19

With this function, given the factors that define a target and a player's social relation, we can predict how many questions this player can answer correctly, that is how much knowledge this player has about the target. The corresponding function coefficients for each category of question are given in Table VII in the appendix.

We compute the R-square value, which is used to evaluate the fit of the regression model. High R-square value is desired in model fitting.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4}$$

In Equation 4, y_i is the sample value, $\hat{y}$ is fitted value and $\overline{y}$ is the average. R-square ($\in [0, 1]$) is the fraction of the variation in the dependent variable (Y) that is explained by the explanatory variables (T, D, C, and E). Our R-square value is 48.3% for all the 24 questions (*Task I*), indicating that our model explains 48.3% of the variability. The results are shown in Table IV.

TABLE IV. REGRESSION STATISTICS

Regression Statistics	All questions	*Sent Subject*	*From Subject*	*Sent Whom*	*From Whom*
Multiple R	0.695	0.551	0.498	0.450	0.454
R Square	0.483	0.304	0.248	0.203	0.206
Standard Error	2.328	0.977	1.100	0.993	1.264
No. questions	1536	384	384	384	384

For *Task II*, the experiment data is further divided by four types of questions, and we run regression analysis between the factors and players' performance under each of the four question categories. The statistics for this fine-grained analysis is in Table IV from columns 3 to 6. These R-squares are lower than 48.3% as expected. This observation is likely due to *i)* smaller data sizes, *ii)* a single question type carries less privacy information.

D. Confidence of Player

For each challenge question, we ask players to enter their confidence about their answers, indicating whether they *know*, *infer*, or *guess* the answer. We correlate the confidence level with player's performance in Table V. We compute the number of correct answers categorized under various confidence levels and types of relations and show the averaged values in the table. The results suggest that an adversary's performance positively correlates to the confidence level. In particular, the ability to correctly infer answers correlates to high performance.

TABLE V. PLAYER'S CONFIDENCE AND PERFORMANCE VS. TYPE OF RELATION

Relation	Within University	Within Department	Undergrad Advisor	Graduate Advisor
	Avg. No. of Correct Answers (Total Selections)			
Guess	4.0 (19.9)	2.8 (16.0)	4.3 (17.5)	4.0 (12.9)
Std. Error	0.3 (1.0)	1.2 (2.2)	0.6 (1.3)	0.5 (1.2)
Infer	1.4 (4.2)	2.8 (8.0)	3.0 (5.9)	4.9 (10.0)
Std. Error	0.4 (1.0)	1.0 (2.2)	0.7 (1.2)	0.6 (1.1)
Know	0 (0)	0 (0)	0.5 (0.7)	1.1 (1.1)
Std. Error	0 (0)	0 (0)	0.2 (0.3)	0.4 (0.4)
Total	5.4 (24)	5.6 (24)	7.8 (24)	9.9 (24)

E. Collusion Simulation

Players may collaborate in the game, which allows them to gain more knowledge about the target. Our collusion analysis is through simulation. The simulation is based on combining the single player performance data. We combine players' answers and select one of their answers as the result of collusion. The selection is based on the confidence of the player for that answer, which we collect for each question during the user study. We choose the answer with the highest player confidence as the collusion result. The result for collusion is in Table VI. The success of collusion increases with the number of players. The data shows that with collusion the number of correct answers increases by up to 3 (out of total 24 questions).

TABLE VI. ADVERSARY COLLUSION ANALYSIS.

Collusion Target	1-man	2-men	3-men
	Avg. No of Correct Answers		
Prof. A	10.8	12.6	14.1
Std. Error	0.5	0.3	0.4
Prof. B	9.8	10.9	12.0
Std. Error	1.2	0.4	0.5
Prof. C	8.0	8.9	8.8
Std. Error	0.6	0.6	0.3
Prof. D	8.1	9.0	9.6
Std. Error	0.9	0.4	0.2

F. Discussion and Summary

The factors we use in the analysis may be correlated with each other, especially for the first 3 factors (T, D, and C). This multicollinearity means that two or more predictor variables (factors) in a multiple regression model are correlated. Multicollinearity only affects the calculations regarding individual predictors. A multiple regression model with correlated predictors can still indicate how well the entire bundle of predictors (all 4 factors) predict the outcome variable Y.

Accurately modeling adversary's knowledge is challenging. Sources of errors in our regression analysis may be due to several reasons, including incompleteness of information, i. e., other sources for a player to gain knowledge are not included in the model. We plan to investigate them in our future work. Nevertheless, this paper describes the first general approach for correlating privacy threat with its sources. We summarize our findings below.

- 41.4% of questions are correctly answered on average, for a student-advisor group where the student plays the role of attacker. This result indicates that a high percentage of email information is guessable by others in the same organization.
- Our regression analysis quantifies the impact of observable factors on information leak. The *duration of relation* and *communication frequency* are two strong predictors.
- We evaluate the impact of collusion among players. The result shows that collusion increases the chance of success in answering the targets' secret questions.
- *Limitations.* The R-square value (48.3%) of our linear regression function is relatively low. Intuitively, the value suggests that our model captures about half of the variations in predicting the adversary's knowledge. Expanding the set of explanatory variables and more sophisticated regression functions are possible ways to improve the prediction accuracy.

Personal information can be leaked via social connections such as work relations. Authentication questions based on personal information could be vulnerable, especially when the adversaries are socially close. We confirms this observation via experimental study in the form of email message based challenge questions. Our approach can be generalized for quantitatively evaluating productive authentication systems that are based on any type of personal or secret information. Such analysis can help make better design choices on appropriate selection and robust presentation of authentication questions.

VI. Conclusions and Future Work

We presented our experimental work on analyzing communication and social factors that contribute to the information gain of the adversary. Our evaluation was conducted in a new role-playing game system with private email data. We used regression analysis to quantify the impact of observable factors on information leak, and found that duration of relation and communication frequency are two strongest predictors of adversary's knowledge in our model.

Our attack game environment provides a means for larger-scale behavior analysis in security paradigm. For future work, we plan to apply behavior-analysis principles to analyze game-players, in particular their decision-making strategies, and to evaluate behavioral-changing interventions that aim at reducing attack activities in players. For example, one may collect and analyze information such as how players calculate the risks and returns in the game, how long players play the games, and how players choose targets and collude. In addition, one can apply behavior-changing triggers such as disincentive/penalty and motivational interventions to the game and compare their effects on players' game-related behaviors. Such studies can provide evidences on the behavioral characteristics and decision-making strategies associated with aggressive cyberspace behaviors.

Another future direction is to expand our model and extending the adversary's knowledge experiments to settings beyond personal privacy, e.g., to model the adversary's knowledge on a system, a server, or an organizational network.

References

[1] Y. Yang, J. Lutes, F. Li, B. Luo, and P. Liu, "Stalking online: on user privacy in social networks," in *Proceedings of the second ACM conference on Data and Application Security and Privacy*, ser. CODASPY '12. New York, NY, USA: ACM, 2012, pp. 37–48. [Online]. Available: http://doi.acm.org/10.1145/2133601.2133607

[2] J. Staddon, "Finding "hidden" connections on LinkedIn an argument for more pragmatic social network privacy," in *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, ser. AISec '09. New York, NY, USA: ACM, 2009, pp. 11–14. [Online]. Available: http://doi.acm.org/10.1145/1654988.1654992

[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: privacy with multidimensional adversarial knowledge," in *Proceedings of the 33rd international conference on Very large data bases*, ser. VLDB '07. VLDB Endowment, 2007, pp. 770–781. [Online]. Available: http://dl.acm.org/citation.cfm?id=1325851.1325939

[4] T. Li, N. Li, and J. Zhang, "Modeling and integrating background knowledge in data anonymization," in *Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on*, 29 2009-april 2 2009, pp. 6 –17.

[5] A. Machanavajjhala, J. Gehrke, and M. Götz, "Data publishing against realistic adversaries," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 790–801, Aug. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1687627.1687717

[6] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge in privacy," in *In ICDE*, 2007, pp. 126–135.

[7] Z. Sun, B. Zan, J. Ban, M. Gruteser, and P. Hao, "Evaluation of privacy preserving algorithms using traffic knowledge based adversary models," in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, oct. 2011, pp. 1075 –1082.

[8] C. Dwork, "Differential privacy," in *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, ser. Lecture Notes in Computer Science, vol. 4052. Venice, Italy: Springer Verlag, July 2006, pp. 1–12. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=64346

[9] F. Dehue, C. Bolman, and T. Vollink, "Cyberbullying: Youngsters' experiences and parental perception," *Cyber Psychology and Behavior*, vol. 11, pp. 217–223, 2008.

[10] M. K. Rogers, K. Seigfried, and K. Tidke, "Self-reported computer criminal behavior: A psychological analysis," *Digit. Investig.*, vol. 3, pp. 116–120, Sep. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.diin.2006.06.002

[11] F. Stajano and P. Wilson, "Understanding scam victims: seven principles for systems security," *Commun. ACM*, vol. 54, no. 3, pp. 70–75, Mar. 2011. [Online]. Available: http://doi.acm.org/10.1145/1897852.1897872

[12] J. Ryoo, A. Techatassanasoontorn, and D. Lee, "Security education using second life," *Computing in Science and Engg.*, vol. 7, no. 2, pp. 71–74, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1109/MSP.2009.49

[13] S. Schechter, A. J. B. Brush, and S. Egelman, "It's no secret. measuring the security and reliability of authentication via "secret" questions," in *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, ser. SP '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 375–390. [Online]. Available: http://dx.doi.org/10.1109/SP.2009.11

[14] M. Smith, C. Szongott, B. Henne, and G. von Voigt, "Big data privacy issues in public social media," in *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, 2012, pp. 1–6.

[15] R. Gross and A. Acquisti, “Information revelation and privacy in online social networks,” in *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, ser. WPES ’05. New York, NY, USA: ACM, 2005, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/1102199.1102214

[16] D. Irani, S. Webb, K. Li, and C. Pu, “Modeling unintended personal-information leakage from multiple online social networks,” *IEEE Internet Computing*, vol. 15, no. 3, pp. 13–19, May 2011. [Online]. Available: http://dx.doi.org/10.1109/MIC.2011.25

[17] M. Srivatsa and M. Hicks, “Deanonymizing mobility traces: using social network as a side-channel,” in *ACM Conference on Computer and Communications Security*, 2012, pp. 628–637.

[18] D. Elsweiler, M. Baillie, and I. Ruthven, “Exploring memory in email refinding,” *ACM Trans. Inf. Syst.*, vol. 26, no. 4, pp. 21:1–21:36, Oct. 2008. [Online]. Available: http://doi.acm.org/10.1145/1402256.1402260

[19] R. M. Gray, *Entropy and information theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1990.

[20] C. E. Shannon, “A mathematical theory of communication,” *Bell Systems Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[21] C. E. Shannon, “Prediction and entropy of printed English,” *Bell Systems Technical Journal*, vol. 30, pp. 50–64, 1951.

[22] H. Wang, K. Yao, G. Pottie, and D. Estrin, “Entropy-based sensor selection heuristic for target localization,” in *Proceedings of the 3rd international symposium on Information processing in sensor networks*, ser. IPSN ’04. New York, NY, USA: ACM, 2004, pp. 36–45. [Online]. Available: http://doi.acm.org/10.1145/984622.984628

[23] C. Ellison, C. Hall, R. Milbert, and B. Schneier, “Protecting secret keys with personal entropy,” *Future Gener. Comput. Syst.*, vol. 16, no. 4, pp. 311–318, Feb. 2000. [Online]. Available: http://dx.doi.org/10.1016/S0167-739X(99)00055-2

[24] M. Jakobsson, L. Yang, and S. Wetzel, “Quantifying the security of preference-based authentication,” in *Proceedings of the 4th ACM workshop on Digital identity management*, ser. DIM ’08. New York, NY, USA: ACM, 2008, pp. 61–70. [Online]. Available: http://doi.acm.org/10.1145/1456424.1456435

[25] Y. Chen and D. Liginlal, “Bayesian networks for knowledge-based authentication,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 5, pp. 695–710, May 2007. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.1024

[26] Y. Chen and D. Liginlal, “A maximum entropy approach to feature selection in knowledge-based authentication,” *Decis. Support Syst.*, vol. 46, no. 1, pp. 388–398, Dec. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2008.07.008

[27] M. Just and D. Aspinall, “Personal choice and challenge questions: A security and usability assessment,” in *Proceedings of the 5th Symposium on Usable Privacy and Security*, ser. SOUPS ’09. New York, NY, USA: ACM, 2009, pp. 8:1–8:11. [Online]. Available: http://doi.acm.org/10.1145/1572532.1572543

[28] D. Balfanz, R. Chow, O. Eisen, M. Jakobsson, S. Kirsch, S. Matsumoto, J. Molina, and P. C. van Oorschot, “The future of authentication,” *IEEE Security & Privacy*, vol. 10, no. 1, pp. 22–27, 2012.

[29] M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang, “Love and authentication,” in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI ’08. New York, NY, USA: ACM, 2008, pp. 197–200. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357087

[30] M. Jakobsson, L. Yang, and S. Wetzel, “Quantifying the security of preference-based authentication,” in *Proceedings of the 4th ACM workshop on Digital identity management*, ser. DIM ’08. New York, NY, USA: ACM, 2008, pp. 61–70. [Online]. Available: http://doi.acm.org/10.1145/1456424.1456435

[31] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano, “The quest to replace passwords: A framework for comparative evaluation of web authentication schemes,” in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, ser. SP ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 553–567. [Online]. Available: http://dx.doi.org/10.1109/SP.2012.44

[32] R. Biddle, S. Chiasson, and P. Van Oorschot, “Graphical passwords: Learning from the first twelve years,” *ACM Comput. Surv.*, vol. 44, no. 4, pp. 19:1–19:41, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2333112.2333114

[33] A. Babic, H. Xiong, D. Yao, and L. Iftode, “Building robust authentication systems with activity-based personal questions,” in *Proceedings of the 2nd ACM workshop on Assurable and usable security configuration*, ser. SafeConfig ’09. New York, NY, USA: ACM, 2009, pp. 19–24. [Online]. Available: http://doi.acm.org/10.1145/1655062.1655067

[34] K. Xu, D. Yao, Q. Ma, and A. Crowell, “Detecting infection onset with behavior-based policies,” in *Network and System Security (NSS), 2011 5th International Conference on*. IEEE, 2011, pp. 57–64.

[35] H. Zhang, D. D. Yao, and N. Ramakrishnan, “Detection of stealthy malware activities with traffic causality and scalable triggering relation discovery,” in *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS ’14. New York, NY, USA: ACM, 2014, pp. 39–50. [Online]. Available: http://doi.acm.org/10.1145/2590296.2590309

[36] K. O. Elish, D. Yao, and B. G. Ryder, “User-centric dependence analysis for identifying malicious mobile apps,” *Workshop on Mobile Security Technologies*, 2012.

Appendix

TABLE VII. Coefficients of the prediction functions for each type of question.

	Sent Subject	From Subject	Sent Whom	From Whom
Intercept	2.70	0.15	-1.10	0.76
Relation Type	0.93	-0.42	-0.26	0.84
Duration	-0.01	0.05	0.03	0.02
Communi-cation	0.05	0.02	0.04	0.02
Entropy	-0.06	0.37	0.58	0.20